# Assignment 5 - CSC/DSC 265/465 - SPRING 2019 - Due May 1

**Q1:** We wish to fit the model

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n, \tag{1}$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independent error terms, and $x_i$ is a predictor variable. The function $g(x)$ has the following properties:

(i) There are three knots $\xi_1 < \xi_2 < \xi_3$.

(ii) $g(x)$ is continuous at the knots.

(iii) $g(x)$ possesses a continuous first derivative at the knots.

(iv) $g(x)$ is a constant $g(x) = a_0$ for $x < \xi_1$.

(v) $g(x)$ is a third order polynomial $g(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$ for $x \in (\xi_1, \xi_2)$.

(vi) $g(x)$ is a third order polynomial $g(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$ for $x \in (\xi_2, \xi_3)$.

(vii) $g(x)$ is a constant $g(x) = a_3$ for $x > \xi_3$.

(a) How many linear constraints are imposed on the parameters $(a_0, a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, a_3)$ by properties (i)-(vii)? Write these explicitly.

(b) Assume the knots $\xi_1, \xi_2, \xi_3$ are known, but the parameters $(a_0, a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, a_3)$ are to be estimated. How many degrees of freedom does this estimation problem possess (that is, how many free parameters are required to completely define $g(x)$)?

**Q2:** For this problem use data set `Cars93` from the `MASS` package. This represents "data from 93 cars on sale in the USA in 1993". This is a data frame which includes the variables `Price` [Midrange Price (in $1,000): average of Min.Price and Max.Price] and `MPG.city` [City MPG (miles per US gallon by EPA rating)]. The object is to determine a functional relationship between *MPG* (response) and *price* (predictor). This will be done using a model of the form

$$y_i = g(x_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $y_i$ is $\log(MPG)$, $x_i$ is $\log(price)$, and $\epsilon_i \sim N(0, \sigma^2)$ The first step, therefore, is to create the transformed variables

```
> log.price = log(Cars93$Price)
> log.mpg = log(Cars93$MPG.city)
```

We consider the following six models:

**M1** $g(x) = \beta_0 + \beta_1 x$, where $\beta_0, \beta_1$ is to be estimated.

**M2** $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, where $\beta_0, \beta_1, \beta_2$ are to be estimated.

**M3** $g(x) = \beta_0 + \beta_1 x^{-1}$, where $\beta_0, \beta_1$ is to be estimated.

**M4** $g(x) = \beta_0 + \beta_1 x^{-2}$, where $\beta_0, \beta_1$ are to be estimated.

**M5** $g(x)$ is a continuous piecewise linear spline with 1 knot at $\xi = 2.5$.

**M6** $g(x)$ is a continuous piecewise linear spline with 2 knots at $\xi = 2.5, 3.0$.

(a) Construct a table which lists, for each model, $SSE$, $AIC$, $BIC$, and $df$ (the model degrees of freedom, including any intercept term, but excluding $\sigma^2$). Use the forms:

$$\begin{aligned} AIC &= n\log(SSE/n) + 2k, \\ BIC &= n\log(SSE/n) + \log(n)k. \end{aligned}$$

where $k$ is the model degrees of freedom.

(b) Repeat the evaluation of $AIC, BIC$, except use the actual log-likelihood:

$$\begin{aligned} AIC &= -2 \times \textit{log-likelihood} + 2k, \\ BIC &= -2 \times \textit{log-likelihood} + \log(n)k. \end{aligned}$$

**HINT:** If `fit` is the output of function `lm`, the log-likelihood for the model can be calculated as in the following example:

```
> fit1 = lm(log.mpg ~ log.price)
> ll1 = logLik(fit1)
```

(c) Recalculate the $AIC$ and $BIC$ scores using the `R` functions `AIC()`, `BIC()`.

(d) For each of the three calculation methods considered above, standardize the $AIC$ and $BIC$ scores by subtracting the minimum:

$$AIC^* = AIC - \min AIC, \quad BIC^* = BIC - \min BIC.$$

Can you conclude that the three model selection procedures are equivalent? (For more on this issue, see Question 4 below).

(e) For each model, create a scatterplot of the data (response on the vertical axis, predictor on the horizontal axis) on which is superimposed the fitted model. Use the predict method to obtain the fitted values using the following grid representing the predictor variable:

```
> xgrid = seq(min(log.price),max(log.price),0.05)
```

Place the six plots on a single graphics window (for example, use command `par(mfrow=c(3,2))`). Add a title to each plot containing the model label M1-M6, and the standardized $AIC^*$, $BIC^*$ values from Part (d) (please round off these values to 2 decimal places for display).

(f) Identify the models with the optimal $AIC$ and $BIC$ scores. Suppose this model selection application is to be based on the $AIC$ score. Examining the plots, is there a distinct reason why the models with the second or third lowest $AIC$ might be used in place of the optimal $AIC$ model?

**Q3:** For this question you will need to install the package `wooldridge` from the `CRAN` package repository (`https://cran.r-project.org/web/packages/`). Using default settings, this can be done using the command `install.packages("wooldridge")`. From this package we will use the data set `lawsch85`:

```
> install.packages("wooldridge")
> library(wooldridge)
> data("lawsch85")
```

The data contains observations of 21 variables for 156 schools. The objective will be to create a predictive model for the response variable `salary` [median starting salary] using 8 other quantitive variables. The response variable will be log-transformed (use the base 10 logarithm `log10()`). To create the data set use the following code:

```
> ### Select variables by column index
>
> vari = c(1,4,5,7,9,15,20,21)
>
> ### Use log-transform for response variable
>
> y = log10(lawsch85$salary)
>
> ### Feature matrix
>
> x = as.matrix(lawsch85[,vari])
>
> ### Create data frame and remove missing values
>
> yx = data.frame(y,x)
> yx = na.omit(yx)
>
> ### Extract separate response and feature matrix
>
> x = as.matrix(yx[,-1])
> y = yx[,1]
>
> ### Names of variables used
>
> names(yx)
[1] "y"       "rank"    "LSAT"    "GPA"     "faculty" "clsize"  "studfac"
[8] "llibvol" "lcost"
>
```

(a) For each of the eight predictors, fit a second order polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

where $y = \log_{10}(\texttt{salary})$ and $x$ is the predictor (you can use the `poly` function within the model formula to generate orthogonal polynomials). Report $R^2$ for each predictor. Which predictor appears to be most informative of salary? Then fit $y$ against all predictors in one model (with no interactions). Plot the residuals against the fitted values. In what way might this model be unsuitable?

(b) Create a new feature matrix $X'$ in the following way. Each predictor contributes two columns to the new feature matrix. These consist of the two columns of orthogonal polynomials created by the `poly()` function with option `degree=2`. Verify that all column variances are equal. Then fit $y$ against all columns of this new feature matrix. Plot the residuals against the fitted values. Compare this plot to the one constructed in Part (a).

(c) Fit a LASSO model using $y$ as response and $X'$ as the feature matrix. Use cross-validation as implemented in `cv.glmnet`. What variables are included in the `1se` solution?

(d) Fit a ridge regression model using $y$ as response and $X'$ as the feature matrix. Use cross-validation as implemented in `cv.glmnet`. Rank the features in order of the absolute value of the coefficients of the `1se` solution of the ridge regression model. What are the ranks of those coefficients selected by the LASSO model of Part (c)?

(e) Compare the variables selected by the `1se` LASSO model to those selected by forward, backwards and all-subset model selection based on the $AIC$ score. Use the `system.time()` function to capture the computation time of each method, then compare these times. The total computation time will be the `elapsed` time. Note that all-subsets regression may take considerably more time to compute. Which features are included in all fitted models (that is, the LASSO model, and all three AIC models). Does this make sense, taking into consideration the $R^2$ values of Part (a)?

(f) Create a scatterplot of `salary` against `rank`. Use the original untransformed values. Superimpose on this plot the fitted values from the LASSO `1se` fitted model, and the best $AIC$ (all subsets) model. You will have to use the untransformed `rank` values, and exponentiate the fitted values $\hat{y}$, that is, plot the values $10^{\hat{y}}$. Use distinct plotting characters and/or colors, and use a legend to indicate the respective sources of the points. Do you see any evidence of a "shrinkage effect"?

**Q4: [For Graduate Students]** This problem is a continuation of Question 2 of this assignment. Show that for a Gaussian multiple regression model the quantities $n\log(SSE/n)$ and $-2 \times$ *log-likelihood* differ by a constant that depends only on $n$, and that therefore, the $AIC$ and $BIC$ evaluation methods of Parts (a) and (b) are equivalent from the point of view of model selection. **HINT:** You may rely on the fact that the MLE of $\sigma^2$ is $SSE/n$, and the MLE of $E[y_i]$ is $\hat{y}_i$.